MATERIALS AND METHODS

DNA Extraction

Total bacterial genomic DNA samples were extracted using the Metagenomic DNA was extracted from all Samples using the PowerMax (stool/soil) DNA isolation kit (MoBio Laboratories, Carlsbad, CA, USA), following the manufacturer's instructions, and stored at −20°C prior to further analysis. The quantity and quality of extracted DNAs were measured using a NanoDrop ND-1000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA,USA) and agarose gel electrophoresis, respectively.

16S rDNA Amplicon Pyrosequencing

PCR amplification of the bacterial 16S rRNA genes V4 region was performed using the forward primer 515F (5'- GTGCCAGCMGCCGCGGTAA -3') and the reverse primer 806R (5'-GGACTACHVGGGTWTCTAAT-3'). Sample-specific paired-end 6-bp barcodes were incorporated into the TrueSeq adaptors for multiplex sequencing. The PCR components contained 25 μl of Phusion High-Fidelity PCR Master Mix, 3 μl (10 uM) of each Forward and Reverse primer, 10 μl of DNA Template, 3μl of DMSO, and 6 μl of ddH 2 O. Thermal cycling consisted of initial denaturation at 98 °C for 30 s, followed by 25 cycles consisting of denaturation at 98 °C for 15 s, annealing at 58 °C for 15 s, and extension at 72 °C for 15 s, with a final extension of 1 min at 72 °C. PCR amplicons were purified with Agencourt AMPure XP Beads (Beckman Coulter, Indianapolis, IN) and quantified using the PicoGreen dsDNA Assay Kit (Invitrogen, Carlsbad, CA, USA). After the individual quantification step, amplicons were pooled in equal amounts, and pair-end 2×150 bp sequencing was performed using the Illlumina NovoSeq6000 platform at GUHE Info technology Co., Ltd (Hangzhou, China).

Sequence Analysis

The Quantitative Insights Into Microbial Ecology (QIIME, v1.9.1) pipeline was employed to process the sequencing data, as previously described (Caporaso, Kuczynski et al. 2010). Briefly, raw sequencing reads with exact matches to the barcodes were assigned to respective samples and identified as valid sequences. The low-quality sequences were filtered through following criteria (Gill, Pop et al. 2006, Chen and Jiang 2014): sequences that had a length of <150 bp, sequences that had average Phred scores of <20, sequences that contained ambiguous bases, and sequences that contained mononucleotide repeats of >8 bp. Paired-end reads were assembled using Vsearch V2.4.4(--fastq_mergepairs --fastq_minovlen 5), than Operational taxonomic unit(OTU) picking using Vsearch v2.4.4, included dereplication(--derep_fulllength),cluster(--cluster_fast,--id 0.97), detectection of chimeras(--uchime_ref) (Rognes 2016). A representative sequence was selected from each OTU using default parameters. OTU taxonomic classification was conducted by VSEARCH searching the representative sequences set against the SILVA132 database (Quast et al. 2013).

An OTU table was further generated to record the abundance of each OTU in each sample and the taxonomy of these OTUs. OTUs containing less than 0.001% of total sequences across all samples were discarded. To minimize the difference of sequencing depth across samples, an averaged, rounded rarefied OTU table was generated by averaging 100 evenly resampled OTU subsets under the 90% of the minimum sequencing depth for further analysis.

Bioinformatics and Statistical Analysis

Sequence data analyses were mainly performed using QIIME and R packages (v3.2.0). OTU-level alpha diversity indices, such as Chao1 richness estimator, ACE metric (Abundance-based Coverage Estimator), PD_whole_tree ,Shannon diversity index, and Simpson index, were calculated using the OTU table in QIIME. OTU-level ranked abundance curves were generated to compare the richness

and evenness of OTUs among samples. Beta diversity analysis was performed to investigate the structural variation of microbial communities across samples using UniFrac distance metrics (Lozupone and Knight 2005, Lozupone, Hamady *et al*. 2007) and visualized via principal coordinate analysis (PCoA), nonmetric multidimensional scaling (NMDS) (Ramette 2007).

Differences in the Unifrac distances for pairwise comparisons among groups were determined using Student's t-test and the Monte Carlo permutation test with 1000 permutations, and visualized through the box-and-whiskers plots. Principal component analysis (PCA) was also conducted based on the genus-level compositional profiles (Ramette 2007). The significance of differentiation of microbiota structure among groups was assessed by PERMANOVA (Permutational multivariate analysis of variance) (McArdle and Anderson 2001) using R package "vegan". Venn diagram was generated to visualize the shared and unique OTUs among samples or groups using R package "VennDiagram", based on the occurrence of OTUs across samples/groups regardless of their relative abundance (Zaura, Keijser *et al*. 2009). Taxa abundances at the phylum, class, order, family, genus and species levels were statistically compared among samples or groups by Kruskal.test from R stats package. LEfSe (Linear discriminant analysis effect size) was performed to detect differentially abundant taxa across groups using the default parameters (Segata, Izard *et al*. 2011). Random forest analysis was applied to discriminating the samples from different groups using the R package "randomForest" with 1,000 trees and all default settings (Breiman 2001, Liaw and Wiener 2002). The generalization error was estimated using 10-fold cross-validation. The expected "baseline" error was also included, which was obtained by a classifier that simply predicts the most common category label. Co-occurrence analysis was performed by calculating Spearman's rank correlations between predominant taxa. Correlations with $|RHO| > 0.6$ and $P < 0.01$ were visualized as co-occurrence network using Cytoscape (Shannon, Markiel *et al*. 2003). Microbial functions were predicted by PICRUSt (Phylogenetic investigation of communities by reconstruction of unobserved states), based on high-quality sequences (Langille, Zaneveld *et al*. 2013). The output file was further analysed using Statistical Analysis of Metagenomic Profiles (STAMP) software package v2.1.3(Parks *et al*., 2014).

Data Access

All raw sequences were deposited in the NCBI Sequence Read Archive under accession number SRP******.

References

Breiman, L. (2001). "Random forests." Machine Learning 45(1): 5-32.

Caporaso, J. G., J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Tumbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld and R. Knight (2010). "QIIME allows analysis of high-throughput community sequencing data." Nature Methods 7(5): 335-336.

Chen, H. and W. Jiang (2014). "Application of high-throughput sequencing in understanding human oral microbiome related with health and disease." Frontiers in Microbiology 5: 6.

Chen, Y. F., F. L. Yang, H. F. Lu, B. H. Wang, Y. B. Chen, D. J. Lei, Y. Z. Wang, B. L. Zhu and L. J. Li (2011)."Characterization of Fecal Microbial Communities in Patients with Liver Cirrhosis." Hepatology 54(2):562-572.

Fadrosh, D.W., Ma, B., Gajer, P., Sengamalay, N., Ott, S., Brotman, R.M. and Ravel. J. (2014) "An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. " Microbiome 2:6.

Gill, S. R., M. Pop, R. T. DeBoy, P. B. Eckburg, P. J. Turnbaugh, B. S. Samuel, J. I. Gordon, D. A. Relman, C. M. Fraser-Liggett and K. E. Nelson (2006). "Metagenomic analysis of the human distal gut microbiome." Science 312(5778): 1355-1359.

Langille, M. G. I., J. Zaneveld, J. G. Caporaso, D. McDonald, D. Knights, J. A. Reyes, J. C. Clemente, D. E. Burkepile, R. L. V. Thurber, R. Knight, R. G. Beiko and C. Huttenhower (2013). "Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences." Nature Biotechnology 31(9): 814-+.

Liaw, A. and M. Wiener (2002). "Classification and regression by randomForest." R News 2(3): 18-22.

Lozupone, C. and R. Knight (2005). "UniFrac: a new phylogenetic method for comparing microbial communities." Applied and environmental microbiology 71(12): 8228-8235.

Lozupone, C. A., M. Hamady, S. T. Kelley and R. Knight (2007). "Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities." Applied and Environmental Microbiology 73(5): 1576-1585.

Parks, D.H., Tyson, G.W., Hugenholtz, P., Beiko, R.G. (2014). "STAMP: Statistical analysis of taxonomic and functional profiles." Bioinformatics 30:3123-3124.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner, F.O. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucl. Acids Res. 41 (D1): D590-D596.

McArdle, B. H. and M. J. Anderson (2001). "Fitting multivariate models to community data: a comment on distance-based redundancy analysis." Ecology 82(1): 290-297.

Ramette, A. (2007). "Multivariate analyses in microbial ecology." Fems Microbiology Ecology 62(2):142-160.

Rognes, T., Flouri, T., Nichols, B., Quince, C., Mahé, C (2016). "VSEARCH: a versatile open source tool for metagenomics. "PeerJ 4: e2584.

Segata, N., J. Izard, L. Waldron, D. Gevers, L. Miropolsky, W. S. Garrett and C. Huttenhower (2011). "Metagenomic biomarker discovery and explanation." Genome Biology 12(6).

Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker (2003). "Cytoscape: A software environment for integrated models of biomolecular interaction networks." Genome Research 13(11): 2498-2504.

Warton, D. I., S. T. Wright and Y. Wang (2012). "Distance-based multivariate analyses confound location and dispersion effects." Methods in Ecology and Evolution 3(1): 89-101.

Zaura, E., B. J. F. Keijser, S. M. Huse and W. Crielaard (2009). "Defining the healthy "core microbiome" of oral microbial communities." Bmc Microbiology 9: 12.